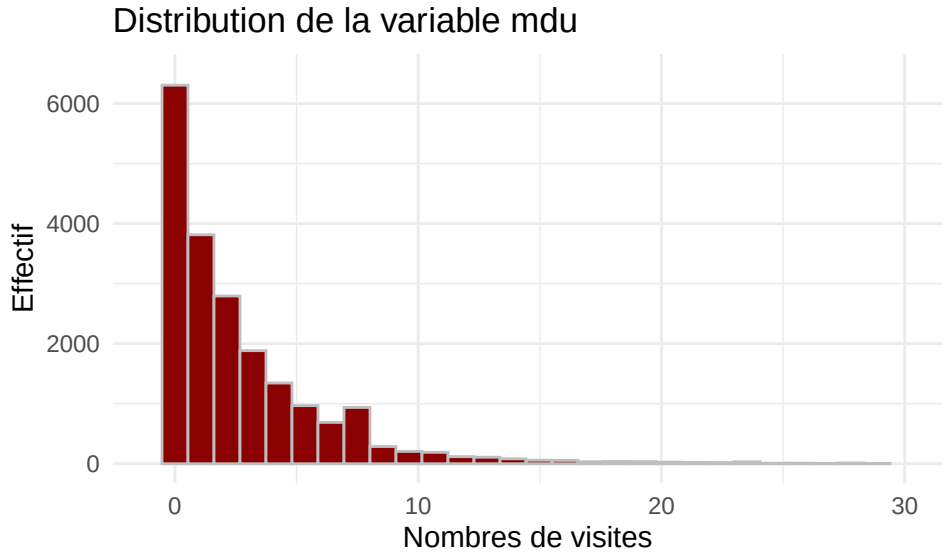
Projet Modèle de comptage M2

Corre_Lenoir

Table des matières

1 Introduction

Nous allons dans ce rapport, estimer le nombre de visites chez le médecin pour un individu. Pour cela, nous allons devoir utiliser des modèles de comptage. Pour appuyer notre travail, nous avons une base de données avec 20184 observations et 19 variables. Notre variable à expliquer sera la variable `mdu` qui correspond au nombre de visite chez le médecin.



La distribution n'est clairement pas Gaussienne-normale mais plutôt de Poisson avec un paramètre λ faible. On remarque une très grande proportion de 0 et l'effectif est décroissant avec le nombre de visites.

Maintenant, voici un résumé des différentes variables qualitatives :

Table 1 – Sommaire variable quantitatives

	mdu	lcoins	lpi	fmde	mdeoff	linc	lfam	age	educdec	ndisease
Min.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1st Qu.	0.00	0.00	4.06	0.00	0.00	8.58	1.10	11.46	11.00	6.90
Median	1.00	3.26	6.11	6.10	450.00	8.98	1.39	24.20	12.00	10.58
Mean	2.86	2.38	4.71	4.03	417.81	8.71	1.25	25.72	11.97	11.24
3rd Qu.	4.00	4.56	6.62	6.96	750.00	9.26	1.61	37.40	13.00	13.73
Max.	77.00	4.56	7.16	8.29	1000.00	10.28	2.64	64.28	25.00	58.60

Pour la variable à expliquer, le nombre de visite varie entre 0 et 77 visites sur la durée de l'étude, avec une médiane à 1, c'est-à-dire que 50% des individus ont été une fois maximum chez le médecin.
Les individus ont un âge moyen de 26 ans avec une taille de famille variant de 1 à 14 personnes.
`lcoins` signifie la part des dépenses de santé qui est payé par le client. Cette part varie entre 0 % et 95 %.

Table 2 – Sommaire des variable qualitatives

	idp	female	black	child	femchild	hlthg	hlthf	hlthp
0	14938	9749	16480	12082	16273	12876	18624	19882
1	5246	10435	3704	8102	3911	7308	1560	302

La base de données est donc composée majoritairement de femmes, de personnes blanches et de personnes de plus de 18 ans.

Pour le niveau de santé 302 personnes se disent en mauvais état de santé, 1560 de santé moyenne et 7308 en bonne santé. Le reste (11014) se disent par défaut en excellente santé.

2 Question 1

Le premier modèle que nous estimerons est le modèle de Poisson simple. Ce modèle repose sur l'hypothèse très forte que l'espérance est égale au paramètre λ qui est égal à la variance. La variable à expliquer `mdu` est une variable de comptage que l'on peut observer que sur $[0; +\infty]$. Premièrement, nous regardons les valeurs de moyenne et de variance :

Table 3 – Moyenne / Variance

	Valeur
moyenne	2.861
variance	20.294

La variance diffère largement de la moyenne. On peut déjà supposer que l'on aura de la sur-dispersion.

Voici le modèle poisson de base que nous estimons :

$$ln(mdu_{it}) = \beta_{0i} + \beta_1 lcoins_{it} + \beta_2 age_{it} + \beta_3 idp_{it} + \beta_4 linc_{it} + \beta_5 female_{it} +$$

$$\beta 6 educdec_{it} + \beta 7 black_{it} + \beta 8 hlthf_{it} + \beta 9 hlthg_{it} + \beta 10 hlthp_{it}$$

Ainsi que les résultats de cette régression :

Table 4 –

	Dependent variable:
	mdu
lcoins	-0.071^{***}
	(0.002)
idp1	-0.129^{***}
	(0.010)
linc	0.070^{***}
	(0.005)
female1	0.291^{***}
	(0.009)
educdec	0.020^{***}
	(0.002)
age	0.004^{***}
	(0.0003)
black1	-0.758^{***}
	(0.015)
hlthg1	0.112^{***}
	(0.009)
hlthf1	0.485^{***}
	(0.015)
hlthp1	0.888^{***}
	(0.026)
Constant	0.109^{**}
	(0.046)
Observations	20,184
Log Likelihood	-61,929.470
Akaike Inf. Crit.	123,880.900
Note:	*p<0.1; **p<0.05; ***p<0.01

On remarque que tous les coefficients sont significatifs.

Pour tester l'importance des variables `lcoins` et `idp`, on doit estimer un modèle sans celles-ci (modèle imbriqué) et faire un test de rapport de vraisemblance :

$$ln(mdu_{it}) = \beta_{0i} + \beta_1 age_{it} + \beta_2 linc_{it} + \beta_3 female_{it} + \beta 4 educdec_{it} + \beta 5 black_{it} + \beta 6 hlthf_{it} + \beta 7 hlthg_{it} + \beta 8 hlthp_{it}$$

Voici les hypothèses du test de rapport de vraisemblance :

H0 : Les modèles complets et imbriqués correspondent tout aussi bien aux données. Par conséquent, on utilise le modèle imbriqué.
H1 : Le modèle complet surpasse considérablement le modèle imbriqué en termes d'ajustement des données. Par conséquent, on utilise le modèle complet.

On obtient la statistique du rapport de vraisemblance de cette manière :

$$stat = 2(L_{Poisson} - L_{Poissoncontraint})$$

Table 5 – Test du rapport de vraisemblance

#Df	LogLik	Df	Chisq	Pr(>Chisq)
11	-61929.47	NA	NA	NA
9	-62739.17	-2	1619.405	0

De ce test, on remarque que la valeur de la statistique du Khi-deux est de 1619.4 et que la p-value est très faible (< au seuil 5%). Donc on rejette H0, l'hypothèse que les deux modèles sont équivalents. Le modèle doit prendre en compte les variables `lcoins` et `idp` puisqu'elles ont un effet sur la régression de Poisson.

3 Question 2

Dans cette question, nous allons évaluer la qualité de l'ajustement. On a remarqué que la variance était supérieure à la moyenne, le modèle de Poisson se basant sur une distribution de poisson ($\lambda = E(Y) = V(Y)$), notre modèle aura surement des problèmes d'ajustement (problème de sur-dispersion)

Table 6 – Valeurs prédites du modèle de Poisson

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
obs	6307	3815	2795	1884	1345	968	689	530	408	287
exp	1971	3712	4228	3707	2718	1743	1009	541	276	137

Les valeurs prédites du modèle de Poisson sont très mal ajustées aux données sauf pour la valeur 1. Quant à elle, la valeur 0 peut poser un gros problème car seulement 31 % sont bien évalués.

3.1 Statistique déviance

On calcule tout d'abord la statistique de deviance grâce aux valeurs estimées du modèle de Poisson. La formule s'écrit :

$$2\sum_{i=1}^{n}(y_i log(\frac{y_i}{expXi\beta}) - (y_i - expX_i\beta))$$

La statistique de deviance est 8.2967031×10^4, maintenant nous testons sa significativité :

La p-value est de 0. Au seuil 5%, les valeurs prédites $\hat{\lambda}$ s'écartent significativement des données observées, le modèle n'est pas bien ajusté.

3.2 Pseudo R2

Vu que le R^2 ne s'interprète pas comme la proportion de variance expliquée par la regression de Poisson, le pseudo R^2 est plus adéquat. Voici sa formule :

$$R^2 = 1 - \frac{D(\hat{\beta})}{D(\hat{\beta}_0)}$$

où $D(\hat{\beta})$ est la déviance résiduelle et $D(\hat{\beta}_0)$ est la déviance du modèle avec seulement la constante.

Le pseudo R^2 est de 0.102. Par rapport au modèle trivial, nous avons une réduction de 10% de la variance. Nous savons que si un modèle est bien ajusté, le R^2 doit être le plus proche possible de 1. Dans notre cas, il est très loin de cette valeur.

3.3 Rapport de vraisemblance

Utile pour tester la significativité globale du modèle. On teste l'écart de déviance.

Voici les hypothèses du test de rapport de vraisemblance :
H_0 : Tous les coefficients sont nuls égaux à 0.
H_1 : Les coefficients sont différents de 0.

Le rapport de vraisemblance a de nombreux point communs avec la statistique de déviance et le pseudo R^2. Voici sa formule :

$$LR = D(\hat{\beta}_0) - D(\hat{\beta})$$

TABLE 7 – Résultat test de significativité du rapport de vraisemblance

	p-value
Valeur	0

La pvalue est très faible, inférieur au seuil 5%. On rejette donc l'hypothèse H_0, le modèle équiprobable n'ajuste pas le modèle. Le modèle est globalement significatif à 5%.

4 Question 3

Comme nous avons vu précédemment, la variance n'est clairement pas égale à la moyenne, et donc cela créer des problèmes de sur-dispersion dans notre cas. Nous allons tester cette sur-dispersion :

- Premièrement, nous prenons les valeurs estimés $\hat{\lambda}$ du modèle. Pour rappel, $\hat{\lambda} = e^{X_i\beta}$
- Ensuite, nous allons tester la sur-dispersion en considérant l'hypothèse suivante : $Var(y_i|X_i) = \lambda_i + \alpha g(\lambda_i)$ où $g(\lambda_i)$ prend soit λ_i soit λ_i^2
- Nous pouvons donc faire une regression MCO sans constante de la forme : $\frac{(y_i - \hat{\lambda_i})^2 - y_i}{\hat{\lambda_i}} = \alpha \frac{g(\hat{\lambda_i})}{\hat{\lambda_i}} + u_i$ où u_i est l'erreur hétéroscédastique.

Nous allons donc tester si :

$H0 : \alpha = 0$ et donc la variance est égale à l'espérance.

$H1 : \alpha > 0$ et donc la variance est différente de l'espérance.

Table 8 –

	Dependent variable:	
	var_cond	
	(1)	(2)
X	5.381***	
	(0.285)	
X2		1.678***
		(0.092)
Observations	20,184	20,184
R^2	0.017	0.016
Adjusted R^2	0.017	0.016
Residual Std. Error (df = 20183)	40.472	40.497
F Statistic (df = 1; 20183)	356.788***	331.304***
Note:	*p<0.1; **p<0.05; ***p<0.01	

On remarque que lorque $g(\lambda_i) = \lambda_i$, la valeur estimée de α est de 5.4 et significative. De même, lorsque $g(\lambda_i) = \lambda_i^2$, la valeur estimée de α est de 1.7 et significative. Les deux α estimés sont positifs, supérieur à 0 et significatifs et donc on est en présence de sur-dispersion puisque la variance n'est pas égale à l'espérance. On conforte notre test avec la commande `dispersiontest` de R et on obtient relativement les mêmes résultats.

Table 9 – Resultat sur-dispersion

	test	**MCO**
alpha	5.381	5.381
alpha^2	1.678	1.678

On remarque bien que alpha est supérieur à 0 significativement et que donc $V[yi|xi]$ est different de λ.

Ici, nous pouvons observer une des limites du modèle de Poisson. Quand une trop grande proportion de zéro est comprise dans la variable de comptage, il peut y avoir de la sur-dispersion, c'est-à-dire qu'il existe une trop forte variation dans les observations. Cela peut entrainer une sous-estimation de la variance des coefficients.

Le modèle de Poisson n'est donc pas le modèle qui s'ajuste le mieux aux donnéees

5 Question 4

Une autre façon d'aborder la sur-dispersion dans le modèle est de modifier notre hypothèse de distribution par le modèle binomial négatif. Ce modèle est une alternative au modèle de Poisson puisqu'il possède un paramètre en plus (celui de dispersion) qui permet d'estimer la variance indépendamment de la moyenne.

Voici les résultats obtenus en comparaison au modèle de Poisson simple:

Table 10 –

	Dependent variable:	
	mdu	
	Poisson	*negative binomial*
	(1)	(2)
lcoins	-0.071^{***}	-0.078^{***}
	(0.002)	(0.005)
idp1	-0.129^{***}	-0.106^{***}
	(0.010)	(0.022)
linc	0.070^{***}	0.074^{***}
	(0.005)	(0.009)
female1	0.291^{***}	0.284^{***}
	(0.009)	(0.018)
educdec	0.020^{***}	0.020^{***}
	(0.002)	(0.003)
age	0.004^{***}	0.004^{***}
	(0.0003)	(0.001)
black1	-0.758^{***}	-0.782^{***}
	(0.015)	(0.027)
hlthg1	0.112^{***}	0.085^{***}
	(0.009)	(0.020)
hlthf1	0.485^{***}	0.471^{***}
	(0.015)	(0.036)
hlthp1	0.888^{***}	0.853^{***}
	(0.026)	(0.074)
Constant	0.109^{**}	0.105
	(0.046)	(0.088)
Observations	20,184	20,184
Log Likelihood	$-61{,}929.470$	$-43{,}175.760$
θ		0.793^{***} (0.011)
Akaike Inf. Crit.	123,880.900	86,373.520
Note:	*p<0.1; **p<0.05; ***p<0.01	

En terme de coefficients, les deux modèles sont relativement égaux mais il est important de montrer que le modèle binomiale négatif est meilleur que une simple modèle de Poisson. On remarque le paramètre $\theta = 0.793$, un paramètre de dispersion, est supérieur à 0 et significatif. Les données sont donc dispersées et on a une préférence pour le modèle binomial négatif.

Nous pouvons également utiliser le rapport de vraisemblance entre les deux modèles (Khi-2 = 3.7509417×10^4). On peut donc dire que l'on rejette l'hypothèse H_0 et donc que le modèle binomial négatif est bien meilleur.

Il existe un autre moyen de prouver la sur-dispersion est de faire la déviance résiduelle sur les degrès de liberté.

Table 11 – Comparaison de sur-dispersion

	D_poisson	**D_bin**
Sur-dispersion	4.113	1.072

La sur-dispersion dans le modèle binomial négatif est beaucoup moins élévée que celle dans le modèle de Poisson. On peut donc dire que le modèle est meilleur. Comme ce rapport est largement supérieur à 1 alors il est préférable de prendre un modèle binomial négatif.

6 Question 5

Avant de commencer à calculer les effets marginaux, regardons quelques chiffres qui pourront nous aider à interpréter les résultats suivant.

TABLE 12 – Pourcentage moyen de co-assurance par rapport à l'état de santé

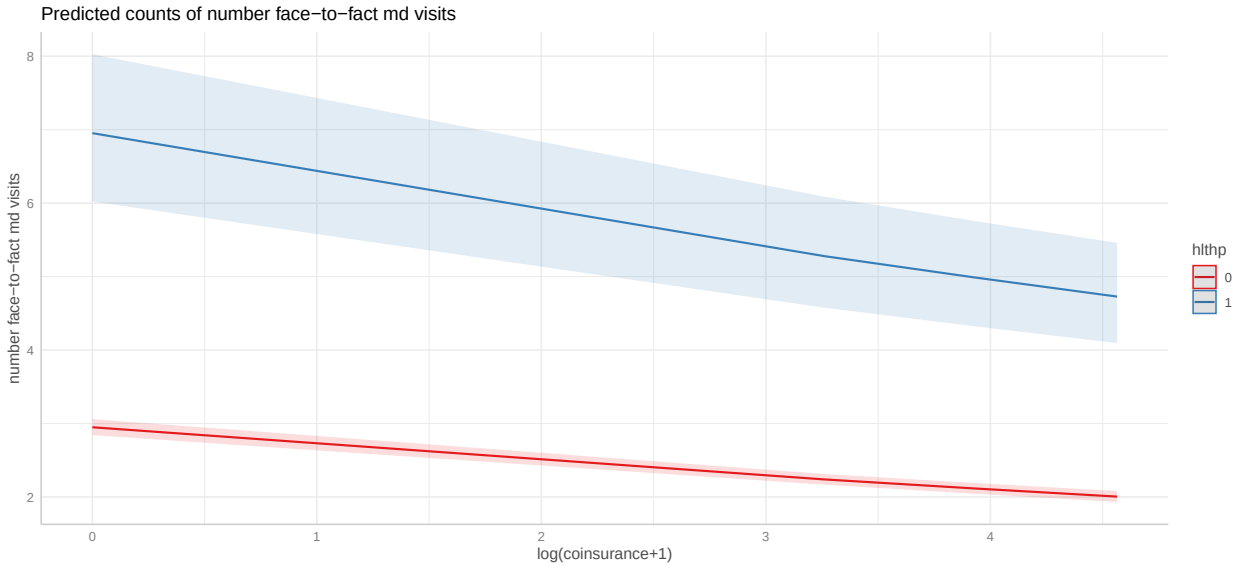
	Mauvais	**Excellente**
Lcoins	8.543	10.92

Ici, on peut voir que le pourcentage moyen des frais médicaux payés par les individus est moins interessant pour les individus qui ont un meilleur état de santé. On peut supposer que les individus se sentant en mauvaise santé payent plus cher une assurance qui rembrousera le plus possible les frais médicaux.

TABLE 13 – Pourcentage moyen de co-assurance par rapport au nombre de visite chez le medecin

	0	**1**	**2**	**10**
Lcoins	16.881	11.741	10.061	5.966

Dans ce tableau, on voit que plus le nombre de visites chez le medecin est important, plus les individus dépensent une part moins importante dans les frais de santé. On peut faire l'hypothèse que plus un individu va chez le médecin, plus il est malade et donc plus il va payer une assurance qui va prendre en charge les frais médicaux.



Ce graphique nous montre bien la même chose que les deux tableaux précédents.

Pour déterminer les effets marginaux, nous appliquons cette formule :

$$\frac{\partial E[y|X]}{\partial x_j} = e^{X\beta}\beta_j$$

où $E[y|X] = e^{X\beta}$

Table 14 – Effets marginaux de lcoins sur mdu

	Lcoins|Excellent	**Lcoins|Mauvais**
Effet marginal	-0.208	-0.435

Ces résultats peuvent s'interpréter de la façon suivante :

- Une variation d'une unité de `lcoins` diminuera les visites chez le medecin de 0,208 unité si l'individu est en excellente santé.
- Pour une unité de `lcoins` en plus, si l'individu est en mauvaise santé, cette variable diminuera les visites chez le medecin de 0.435 unité. Donc plus un individu s'estime en mauvaise santé, pour une variation de `lcoins` l'impactera.

Une autre façon de déterminer les effets marginaux directement par R est la commande `ngbinmfx`.

7 Question 6

Danc cette question, nous allons réalisé un modèle Hurdle. En effet, nous avons énormément de zéros et dans un modèle poisson, il se peut que les valeurs nulles proviennent d'un effet individuel d'aller ou non chez le médecin. Le modèle Hurdle va tenter de supprimer cela. Le modèle se divise en deux parties : la première va regarder si l'individu va chez le médecin ou non (estimé par un logit binomial) qui est le "zero hurdle model" et une deuxième qui va estimer les valeurs positives, combien de fois l'individu va chez le medecin (estimé par un poisson binomial négatif tronqué) qui est le "count model". Voici les estimations du modèle :

Table 16 – Estimations du modèle Hurdle

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
obs	6307	3815	2795	1884	1345	968	689	530	408	287
exp	6307	4101	2543	1744	1255	929	702	537	416	325

En terme de prédiction le modèle hurdle est vraiment très efficace comparé au modèle de poisson simple.

En ce qui concerne la comparaison des deux modèles (Voir table ci-dessous), on se rend compte dans la partie qui estime les zéros, tous les coefficents sont significatifs comme le modèle de poisson mais dans la partie modèle de comptage, les variables `idp` et `educdec` ne sont plus significatives.

Interprétation des coefficients du modèle Hurdle :

Pour interpréter l'effet des coefficients, nous faisons : $exp(\hat{\beta})$
En ce qui concerne le modèle Zéro-Hurdle : Nous avons 1.68 plus de chances d'aller au moins 1 fois chez le medecin si on est une femme (coefficient positif et significatif) et 2.76 fois plus de chance si l'on s'estime en mauvaise santé mais au contraire on a 0.27 moins de chance d'aller chez le médecin si l'individu est noir (coefficient négatif et significatif). Plus l'état de santé estimé de l'individu se dégrade, plus il a de chances d'aller chez le médecin.

Pour le modèle avec les valeurs positives, nous avons un nombre moyen de 1.4 visites (individu en excellente santé, homme, blanc) qui augmente de 1.2 chances si l'individu est une femme, de 2.32 si la personne s'estime en mauvaise santé et baisse de 0.67 fois si l'individu est noir.

En se basant clairement sur les prédictions et le fait que les coefficients sont de mêmes signes et significatifs pour Poisson et Hurlde, d'après nous, **le modèle Hurdle est le modèle que nous devons conserver et qui explique le mieux les données**. De plus, on obtient comme valeur du vraisemblance 3.8415088×10^4 donc le modèle Hurdle est meilleur que le modèle de Poisson.

TABLE 15 –

	Dependent variable: mdu		
	Hurdle		
	count	zero-hurdle	Poisson
lcoins	-0.052^{***}	-0.137^{***}	-0.071^{***}
	(0.006)	(0.008)	(0.002)
idp1	-0.036	-0.268^{***}	-0.129^{***}
	(0.029)	(0.037)	(0.010)
linc	0.037^{**}	0.106^{***}	0.070^{***}
	(0.012)	(0.014)	(0.005)
female1	0.186^{***}	0.520^{***}	0.291^{***}
	(0.024)	(0.032)	(0.009)
educdec	0.006	0.055^{***}	0.020^{***}
	(0.005)	(0.006)	(0.002)
age	0.004^{***}	0.002^{***}	0.004^{***}
	(0.001)	(0.001)	(0.0003)
black1	-0.40^{***}	-1.29^{***}	-0.758^{***}
	(0.041)	(0.042)	0.015
hlthg1	0.103^{***}	0.093^{***}	0.112^{***}
	(0.027)	(0.036)	(0.009)
hlthf1	0.484^{***}	0.426^{***}	0.485^{***}
	(0.049)	(0.067)	(0.015)
hlthp1	0.841^{***}	1.017^{***}	0.888^{***}
	(0.094)	(0.157)	(0.026)
Constant	0.341^{**}	-0.514^{***}	0.109^{**}
	(0.120)	(0.143)	(0.046)
$\ln(\theta)$	-0.642^{***}		
	(0.121)		
Observations	20,184		20,184
Log Likelihood	$-42{,}715.220$		-61,929.470
Note:			*p<0.1; **p<0.05; ***p<0.01